National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

High Performance Open Source Platform for Ocean Sciences

Thomas Huang
Data Scientist | Principal Investigator | Technologist | Architect
thomas.huang@jpl.nasa.gov

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsements by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Data Scientist @NASA/JPL

Principal Investigator
NASA AIST OceanWorks – Ocean Science Platform on Cloud

Project Technologist
NASA’s Physical Oceanography Distributed Active Archive Center (PO.DAAC)

Co-Investigator and Architect
NASA Sea Level Change Portal

Architect
CEOS Ocean Variables Enabling Research and Application for GEOS (COVERAGE)

Architect
Tactical Data Science Framework for Naval Research

Cluster Chair
Federation of Earth Science Information Partners (ESIP) Cloud Computing

Previously Principal Investigator / Co-Investigator
Several NASA-funded Big Data Analytic Projects – Big Data Analytics on the Cloud, Anomaly Detection, In Situ and Satellite Matchup, Search Relevancy, and Quality Screening

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NASA Sea Level Change Portal – *https://sealevel.nasa.gov*

PI: Dr. Carmen Boening, JPL

Goal for the NASA Sea Level Change Team

- Determine how much will sea level rise by [2100]?
- What are the key sensitivities?
- Where are the key uncertainties? Observables? Model Improvements

Goals for the NASA Sea Level Change Portal

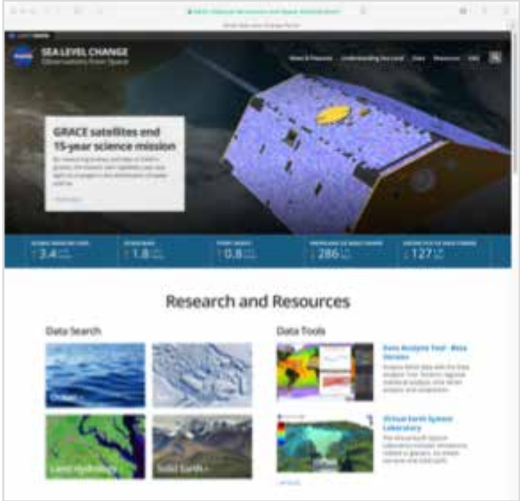
- Provide scientists and the general public with a “one-stop” source for current sea level change information and data
- Provide interactive tools for analyzing and viewing regional data
- Provide virtual dashboard for sea level indicators
- Provide latest news, quarterly report, and publications
- Provide ongoing updates through a suite of editorial products

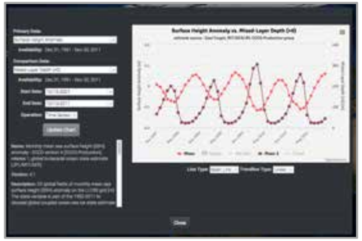
Requires

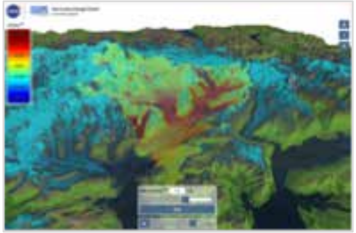
- Interdisciplinary collaboration
- Connect disciplines and evaluate dependencies

Sea Level Change Portal facilitates

- Easy interdisciplinary data comparison
- Access to latest news and information
- Collaboration (data and information exchange)







Web, Social Medias, and Headliners

- 373K monthly page views
- 172K sessions
- 143K users
- **Social Medias**

  Twitter: @NASASeaLevel has over 23K followers

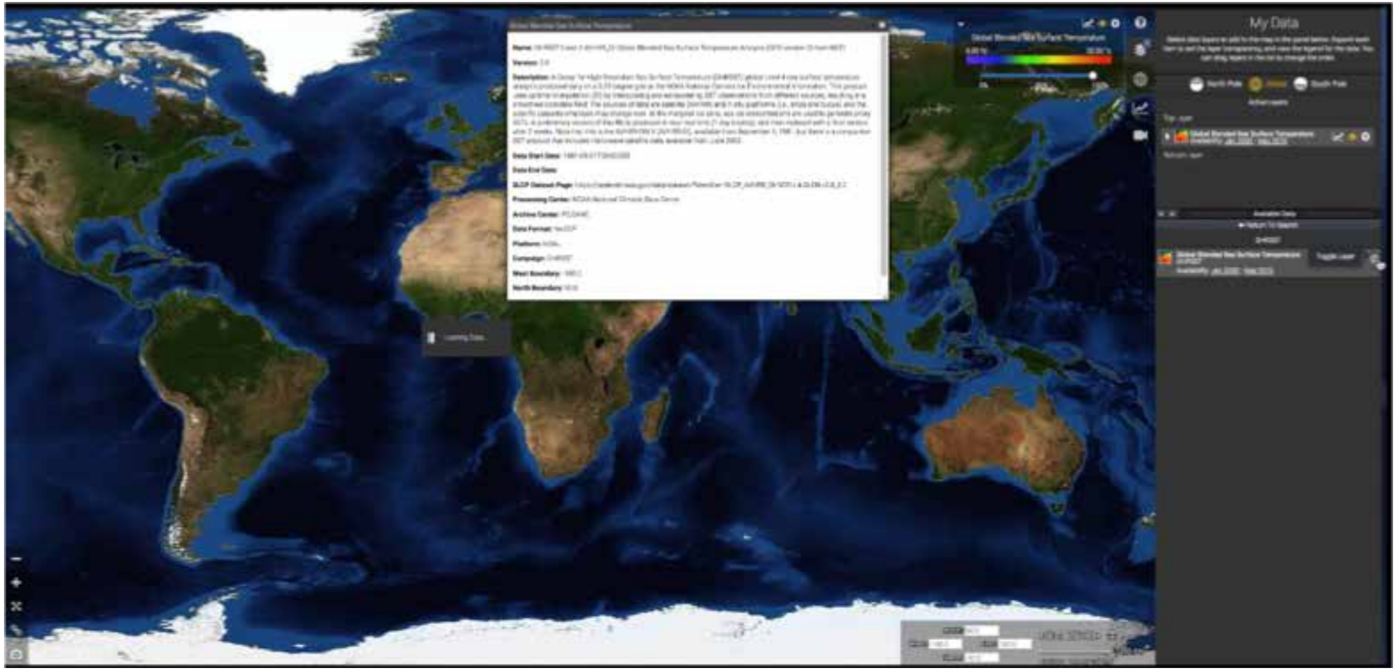
  Facebook: over 31K followers

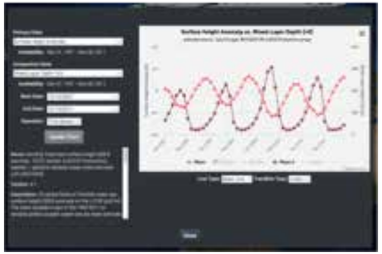
<u>TECH HEADLINES</u>

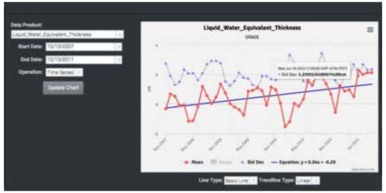
“NASA Sea Level Change Website Offers Everything You Need to Know About Climate Change”
http://www.techtimes.com/articles/147210/20160405/nasa-sea-level-change-website-offers-everything-need-know-climate.htm

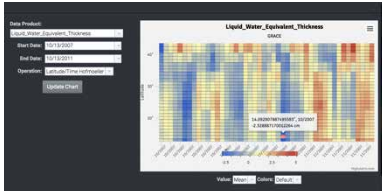
“NASA’s New Sea Level Site Puts Climate Change Papers, Data, and Tools Online”
http://techcrunch.com/2016/04/04/nasas-new-sea-level-site-puts-climate-change-papers-data-and-tools-online/

Analyze Sea Level On-The-Fly *https://sealevel.nasa.gov*






Data Product: Liquid_Water_Equivalent_Thickness
Start Date: 10/13/2007
End Date: 10/13/2011
Operation: Time Series
Update Chart
Liquid_Water_Equivalent_Thickness
GRACE
Mean
Std Dev
Equation: y = 0.04x + -0.29
Line Type: Basic Line
Trendline Type: Linear


Data Product: Liquid_Water_Equivalent_Thickness
Start Date: 10/13/2007
End Date: 10/13/2011
Operation: Latitude/Time Hofmoeller

Value: Mean
Colors: Default

Sea Level Change - Data Analysis Tool

Visualizations | Hydrological Basins | Time Series | Deseason | Data Comparison | Scatter Plot | Latitude/Time Hovmöller | Etc.

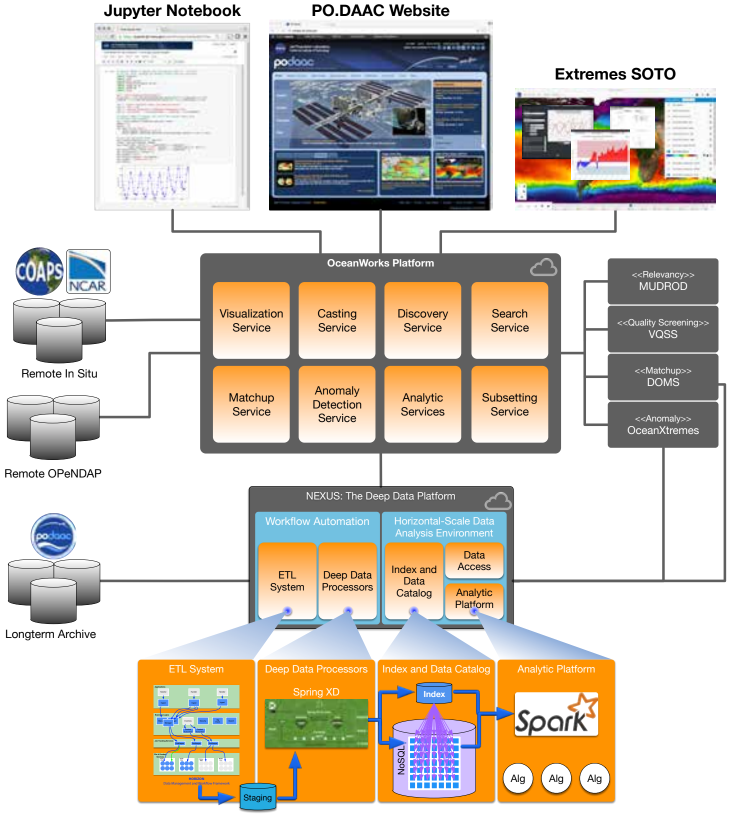
Big Data and Data Centers

- **Increasing "big data" era is driving needs to**
  - Scale computational and data infrastructures
  - Support new methods for deriving scientific inferences
  - Shift towards integrated data analytics
  - Apply computation and data science across the lifecycle
- **For NASA Data Centers, with large amount of observational and modeling data, downloading to local machine is becoming inefficient**
- **Reality with large amount of observational and modeling data**
  - Downloading to local machine is becoming inefficient
  - Search has gotten a lot faster. Too many matches
  - Finding the relevant measurement has becoming a very time consuming process "*Which SST dataset I should use?*"
  - Analyze decades of regional measurement is labor-intensive and costly
- **Limitations**
  - Little to no interoperability between tools and services: metadata standard, keyword, spatial coverage (0-360 or -180..180), temporal representation, etc.
  - Making sure the most relevant measurements return first
  - Visualization is nice, but it doesn't provide enough information about the event/phenomenon captured in the image.
  - With large amount of observational data, data centers need to do more than just storing bits

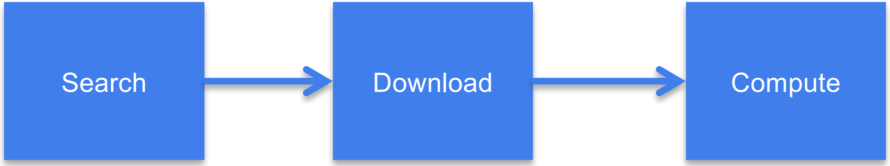
AIST OceanWorks

PI: Thomas Huang

- **OceanWorks** is to establish an **Integrated Data Analytic Center** at the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) for Big Ocean Science
- Focuses on technology integration, advancement and maturity
- Collaboration between JPL, FSU, NCAR, and GMU
- Bringing together PO.DAAC-related big data technologies
  - Anomaly detection and ocean science
  - Big data analytic platform
  - Distributed in-situ to satellite matchup
  - Search relevancy and discovery – linking datasets, services, and anomalies through recommendations
  - Metadata translation and services aggregation
  - Fast data subsetting
  - Virtualized Quality Screening Service

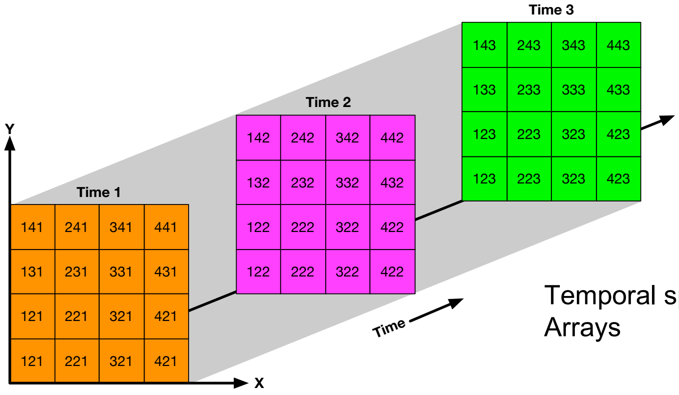


Traditional Method for Analyze Satellite Measurements



Search → Download → Compute

- Depending on the data volume (size and number of files)
- It could take many hours of download – (e.g. 10yr of observational data could yield thousands of files)
- It could take many hours of computation
- It requires expensive local computing resource (CPU + RAM + Storage)
- After result is produced, purge downloaded files



Time 1

141	241	341	441
131	231	331	431
121	221	321	421
121	221	321	421

Time 2

142	242	342	442
132	232	332	432
122	222	322	422
122	222	322	422

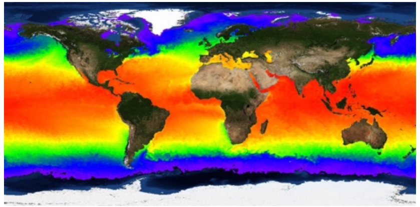
Time 3

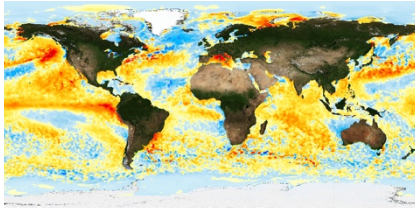
143	243	343	443
133	233	333	433
123	223	323	423
123	223	323	423

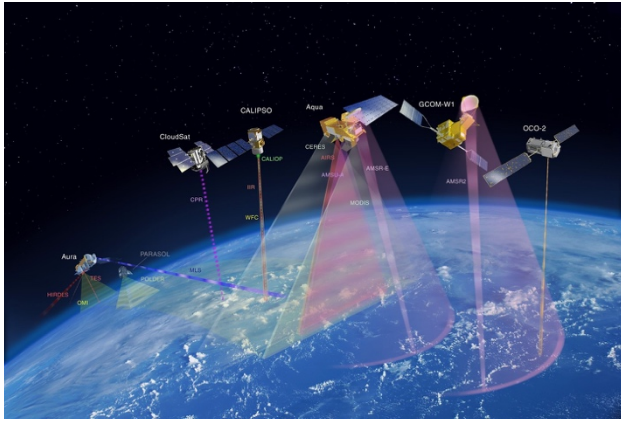
Temporal spatial Arrays

Observation

- Traditional methods for data analysis (time-series, distribution, climatology generation) can't scale to handle large volume, high-resolution data. They perform poorly
- Performance suffers when involve large files and/or large collection of files
- A high-performance data analysis solution must be free from file I/O bottleneck

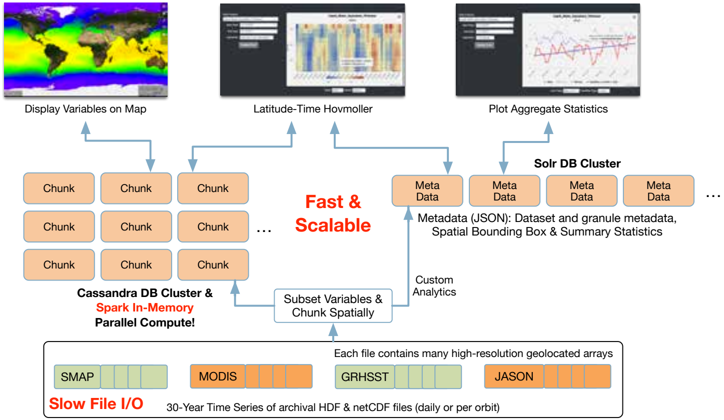




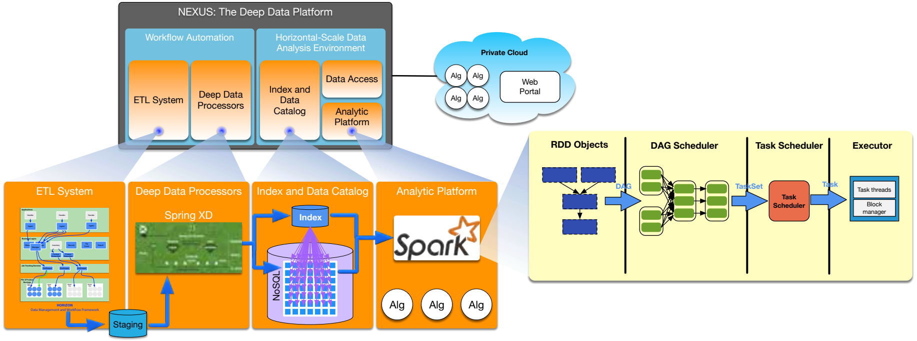


NEXUS: Scalable Data Analytic Solution

- NEXUS is a data-intensive analysis solution using a new approach for handling science data to enable large-scale data analysis
- Streaming architecture for horizontal scale data ingestion
- Scales horizontally to handle massive amount of data in parallel
- Provides high-performance geospatial and indexed search solution
- Provides tiled data storage architecture to eliminate file I/O overhead
- A growing collection of science analysis webservices using Apache Spark: parallel compute, in-memory map-reduce framework
- Pre-Chunk and Summarize Key Variables
    - Easy statistics instantly (milliseconds)
    - Harder statistics on-demand using Spark (in seconds)
    - Visualize original data (layers) on a map quickly (Cassandra store)
- **Algorithms** – Time Series | Latitude/Time Hovmöller| Longitude/Time Hovmöller| Latitude/Longitude Time Average | Area Averaged Time Series | Time Averaged Map | Climatological Map | Correlation Map | Daily Difference Average



Two-Database Architecture



Open Source: Apache License 2
https://github.com/apache/incubator-sdap-nexus

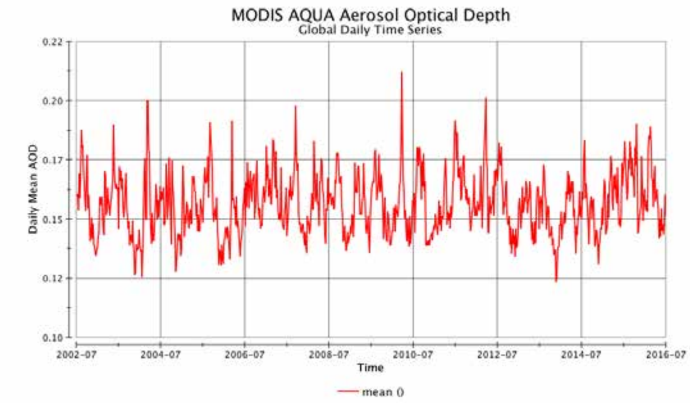
NEXUS Performance: GIOVANNI vs. Custom Spark vs. AWS EMR

Dataset: MODIS AQUA Daily
Name: Aerosol Optical Depth 550 nm (Dark Target) (MYD08_D3v6)
File Count: 5106
Volume: 2.6GB
Time Coverage: July 4, 2002 – July 3, 2016

Giovanni: A web-based application for visualize, analyze, and access vast amounts of Earth science remote sensing data without having to download the data.

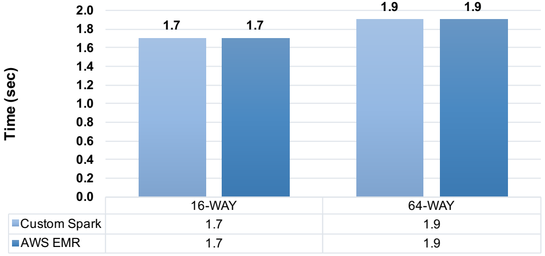
- Represents current state of data analysis technology, by processing one file at a time
- Backed by the popular NCO library. Highly optimized C/C++ library

AWS EMR: Amazon’s provisioned MapReduce cluster



Area Averaged Time Series on AWS - Boulder
July 4, 2002 - July 3, 2016
NEXUS Performance

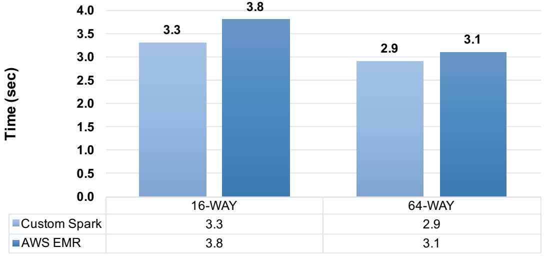
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1140.22 sec



	16-WAY	64-WAY
Custom Spark	1.7	1.9
AWS EMR	1.7	1.9

Area Averaged Time Series on AWS - Colorado
July 4, 2002 - July 3, 2016
NEXUS Performance

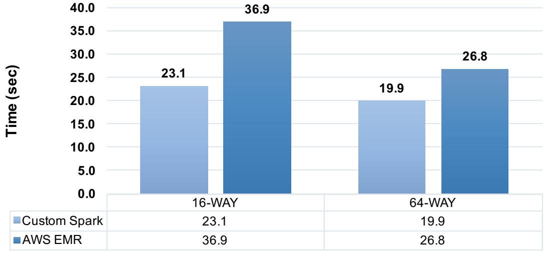
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1150.6 sec



	16-WAY	64-WAY
Custom Spark	3.3	2.9
AWS EMR	3.8	3.1

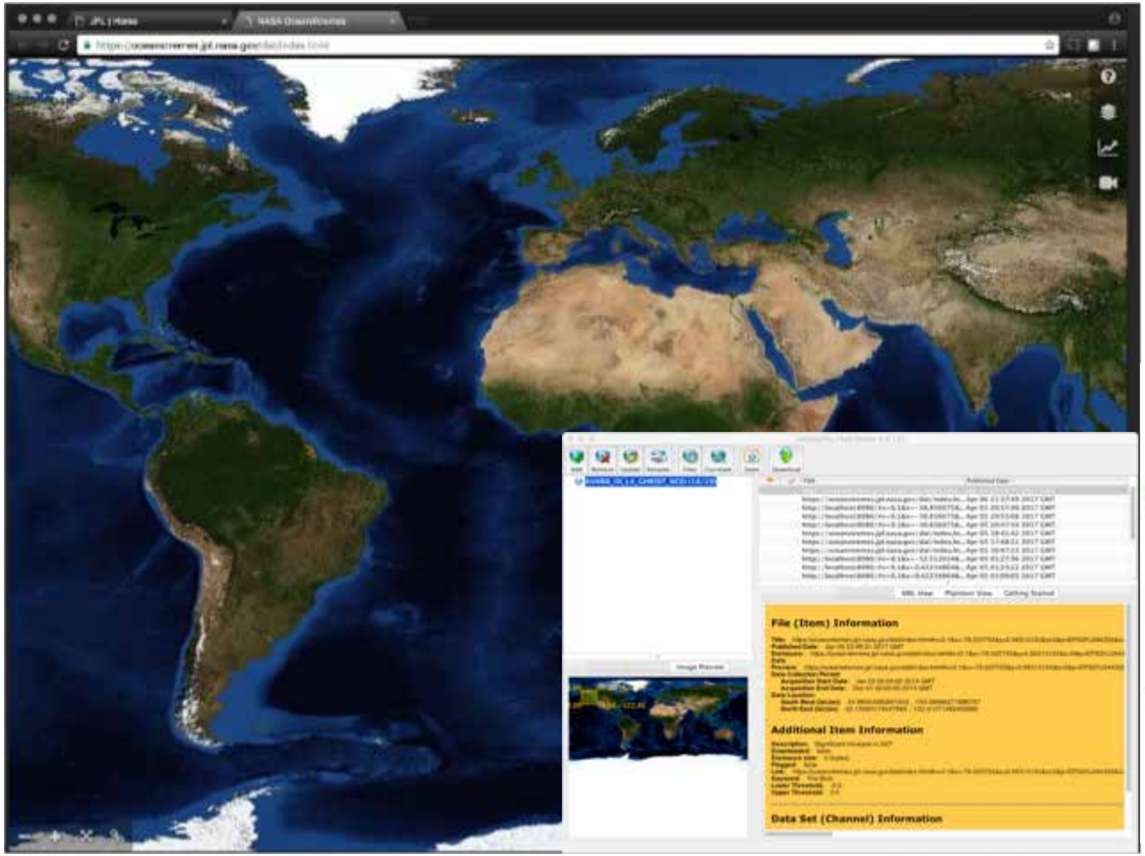
Area Averaged Time Series on AWS - Global
July 4, 2002 - July 3, 2016
NEXUS Performance

Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1366.84 sec



	16-WAY	64-WAY
Custom Spark	23.1	19.9
AWS EMR	36.9	26.8

Analyze Ocean Anomaly – “The Blob”



File (Item) Information

Additional Item Information

Data Set (Channel) Information

- **Visualize** parameter
- **Compute** daily differences against climatology
- **Analyze** time series area averaged differences
- **Replay** the anomaly and visualize with other measurements
- **Document** the anomaly
- **Publish** the anomaly

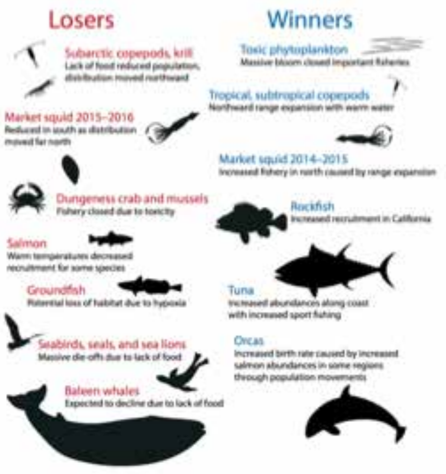
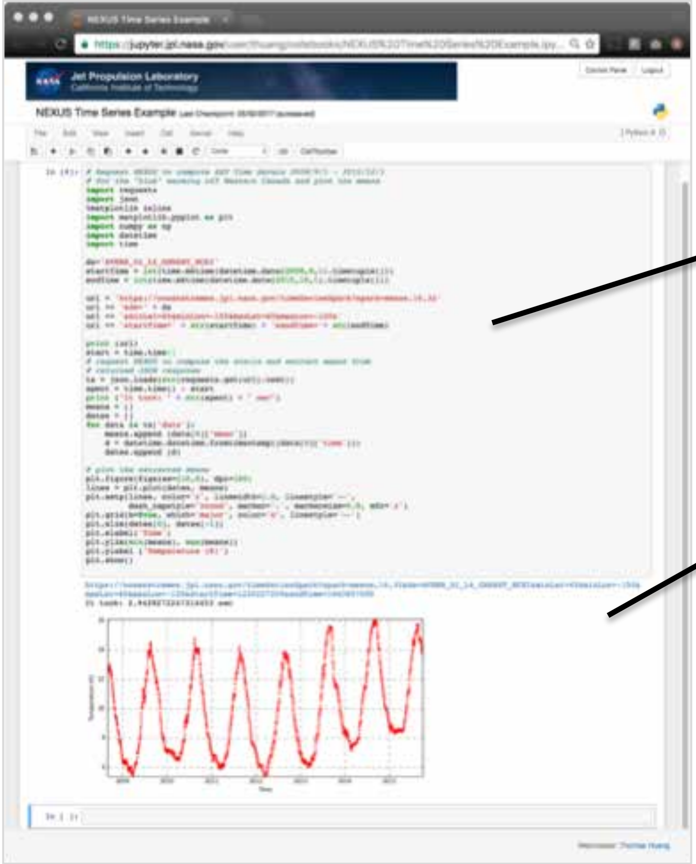

Losers
Subarctic copepods, krill
Lack of food reduced population, distribution moved northward
Market squid 2015–2016
Reduced in south as distribution moved far north
Dungeness crab and mussels
Fishery closed due to toxicity
Salmon
Warm temperatures decreased recruitment for some species
Groundfish
Potential loss of habitat due to hypoxia
Seabirds, seals, and sea lions
Massive die-offs due to lack of food
Baleen whales
Expected to decline due to lack of food
Winners
Toxic phytoplankton
Massive bloom closed important fisheries
Tropical, subtropical copepods
Northward range expansion with warm water
Market squid 2014–2015
Increased fishery in north caused by range expansion
Rockfish
Increased recruitment in California
Tuna
Increased abundances along coast with increased sport fishing
Orcas
Increased birth rate caused by increased salmon abundances in some regions through population movements

Figure from Cavole, L. M., et al. (2016). "Biological Impacts of the 2013–2015 Warm-Water Anomaly in the Northeast Pacific: Winners, Losers, and the Future." Oceanography 29.

Enable Science without File Download


NEXUS Time Series Example
Jet Propulsion Laboratory
California Institute of Technology
NEXUS Time Series Example

```
# Request NEXUS to compute SST Time Series 2008/9/1 - 2015/10/1
# for the "blob" warming off Western Canada and plot the means
…
ds='AVHRR_OI_L4_GHRSST_NCEI'

url = … # construct the webservice URL request


# make request to NEXUS using URL request
# save JSON response in local variable
ts = json.loads(str(requests.get(url).text))

# extract dates and means from the response
means = []
dates = []
for data in ts['data']:
    means.append (data[0]['mean'])
    d = datetime.datetime.fromtimestamp((data[0]['time']))
    dates.append (d)

# plot the result
…
```

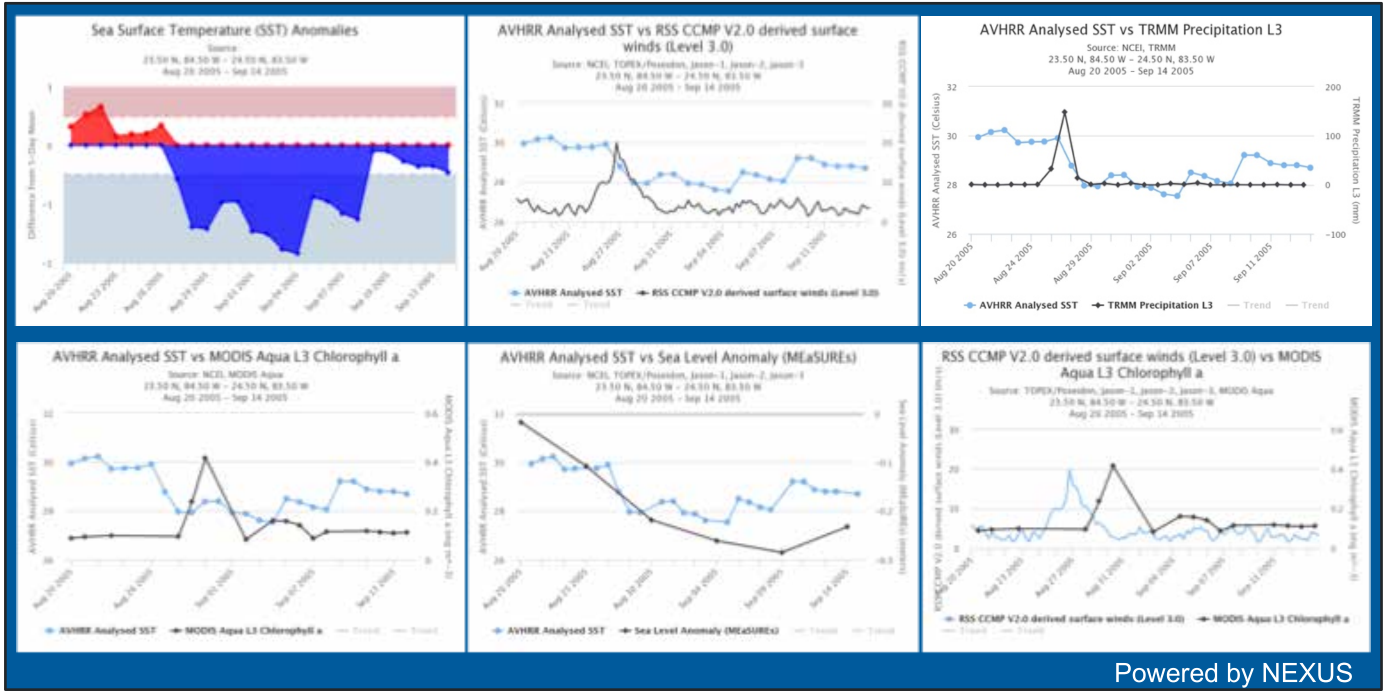
```
https://oceanxtremes.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GHRSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=1220227200&endTime=1443657600

It took: 2.9428272247314453 sec
```

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Hurricane Katrina Study


Sea Surface Temperature (SST) Anomalies

AVHRR Analysed SST vs RSS CCMP V2.0 derived surface winds (Level 3.0)

AVHRR Analysed SST vs TRMM Precipitation L3

Source: NCEI, TRMM
23.50 N, 84.50 W - 24.50 N, 83.50 W
Aug 20 2005 - Sep 14 2005

AVHRR Analysed SST vs MODIS Aqua L3 Chlorophyll a

AVHRR Analysed SST vs Sea Level Anomaly (MEaSUREs)

RSS CCMP V2.0 derived surface winds (Level 3.0) vs MODIS Aqua L3 Chlorophyll a

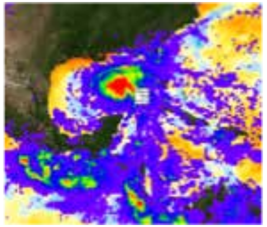
Powered by NEXUS

Source: NCEI, TOPEX/Poseidon, Jason-1, Jason-2, Jason-3
Source: NCEI, MODIS Aqua
Source: TOPEX/Poseidon, Jason-1, Jason-2, Jason-3, MODIS Aqua
23.50 N, 84.50 W - 24.50 N, 83.50 W
Aug 20 2005 - Sep 14 2005
Difference from 5-Day Mean
AVHRR Analysed SST (Celsius)
TRMM Precipitation L3 (mm)
MODIS Aqua L3 Chlorophyll a (mg m^-3)
Sea Level Anomaly (MEaSUREs) (meters)
32
30
28
26
200
100
0
−100
Aug 20 2005
Aug 24 2005
Aug 29 2005
Sep 02 2005
Sep 07 2005
Sep 11 2005
AVHRR Analysed SST
RSS CCMP V2.0 derived surface winds (Level 3.0)
TRMM Precipitation L3
MODIS Aqua L3 Chlorophyll a
Sea Level Anomaly (MEaSUREs)
Trend
Trend

Hurricane Katrina passed to the southwest of Florida on Aug 27, 2005. The ocean response in a 1 x 1 deg region is captured by a number of satellites. The initial ocean response was an immediate cooling of the surface waters by 2 ℃ that lingers for several days. Following this was a short intense ocean chlorophyll bloom a few days later. The ocean may have been “preconditioned’ by a cool core eddy and low sea surface height.

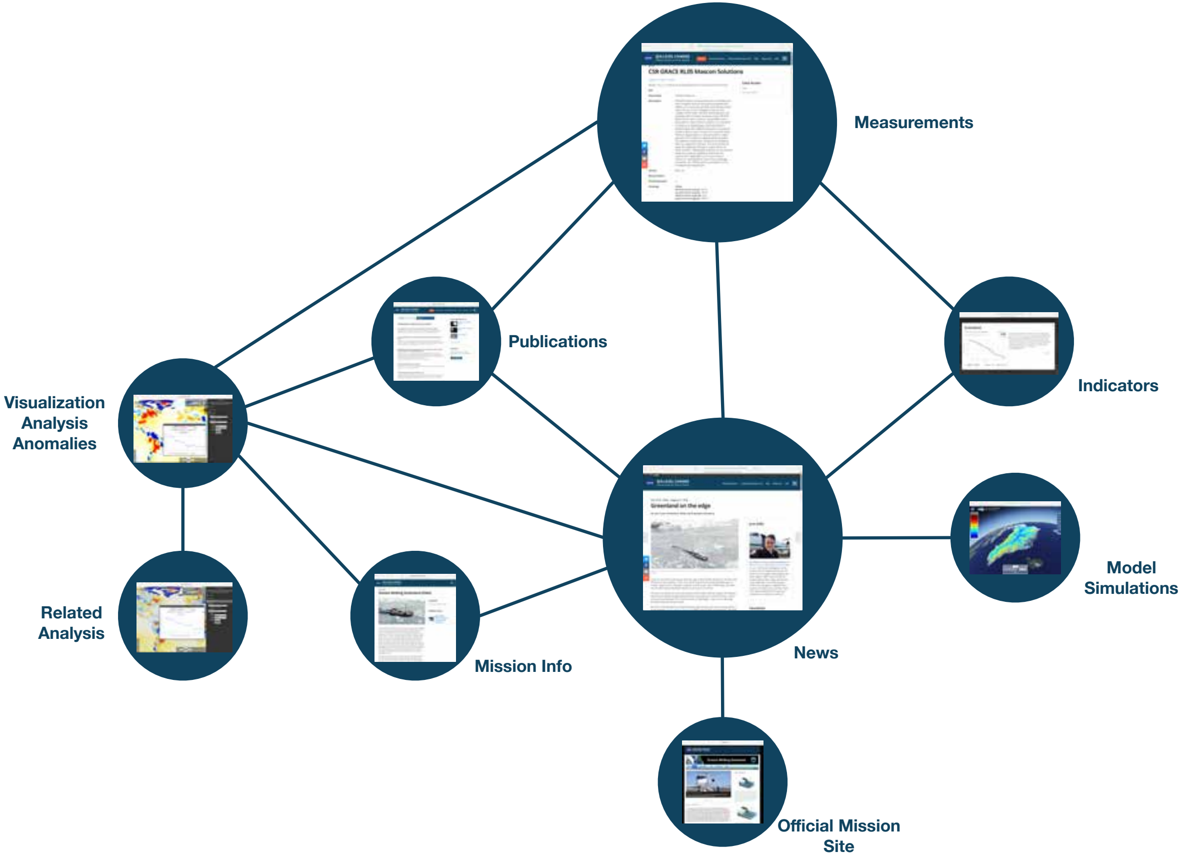
The SST drop is correlated to both wind and precipitation data. The Chl-A data is lagged by about 3 days to the other observations like SST, wind and precipitation.

A study of a Hurricane Katrina–induced phytoplankton bloom using satellite observations and model simulations
Xiaoming Liu, Menghua Wang, and Wei Shi
JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 114, C03023, doi:10.1029/2008JC004934, 2009



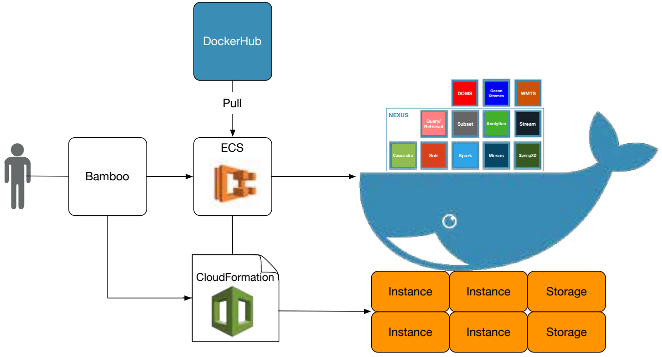
Hurricane Katrina TRMM overlay SST Anomaly

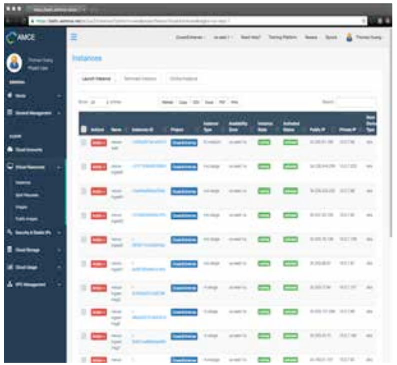
Developing Information Discovery Solutions



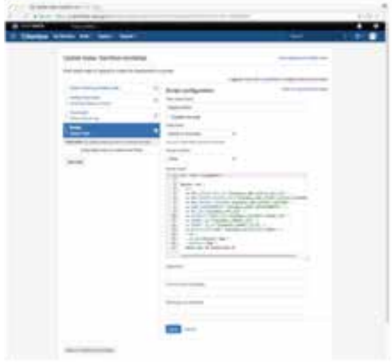
Deployment Automation

- Cloud Deployment is nontrivial
- Infrastructure Definition
  - Various machine instances
  - Storage and buckets
- Software Deployment.. manually
  - Build
  - Package
  - Install
  - Configure
  - Shell login (security issues)
- Best Practice: Deployment Automation
  - Script Infrastructure Definition (e.g. Amazon CloudFormation)
  - Container-based Deployment (e.g. Amazon ECS and DockerHub)





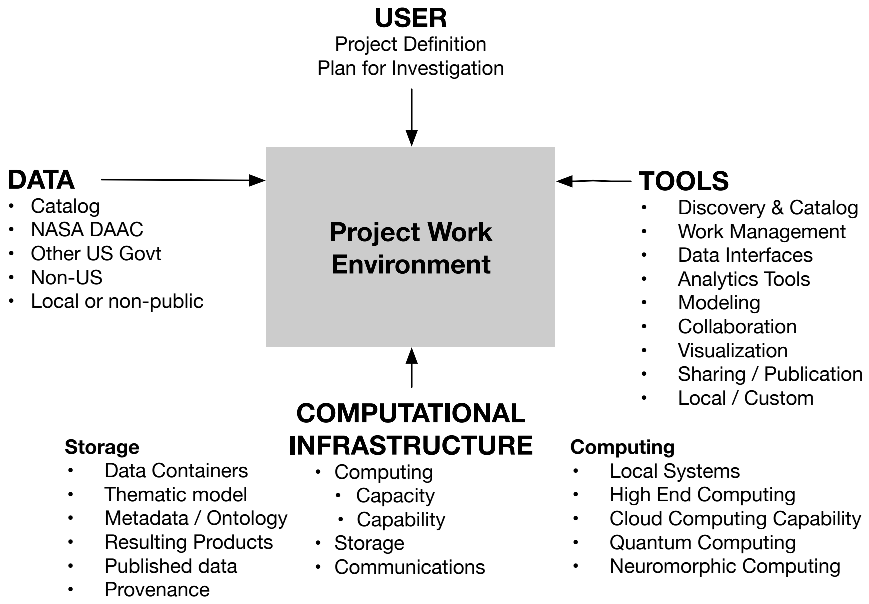
AMCE Deployment



NGAP Deployment

Integrated Data Analytic Center

- An environment for conducting a Science investigation
  - Enables the confluence of resources for that investigation
  - Tailored to the individual study area (ocean, atmospheric, sea level, etc.)
- Harmonizes data, tools and computational resources to permit the research community to focus on the investigation
  - Reduce the data preparation time to something tolerable
  - Catalog of optional resources
  - Semantic-enabled catalog of resources
  - Relevant publications
  - Provide established training data sets of varying resolution
  - Provide effective project confidentiality, integrity and availability
  - Single sign-on and unified financial tracking



USER
Project Definition
Plan for Investigation

Project Work Environment

DATA
- Catalog
- NASA DAAC
- Other US Govt
- Non-US
- Local or non-public

TOOLS
- Discovery & Catalog
- Work Management
- Data Interfaces
- Analytics Tools
- Modeling
- Collaboration
- Visualization
- Sharing / Publication
- Local / Custom

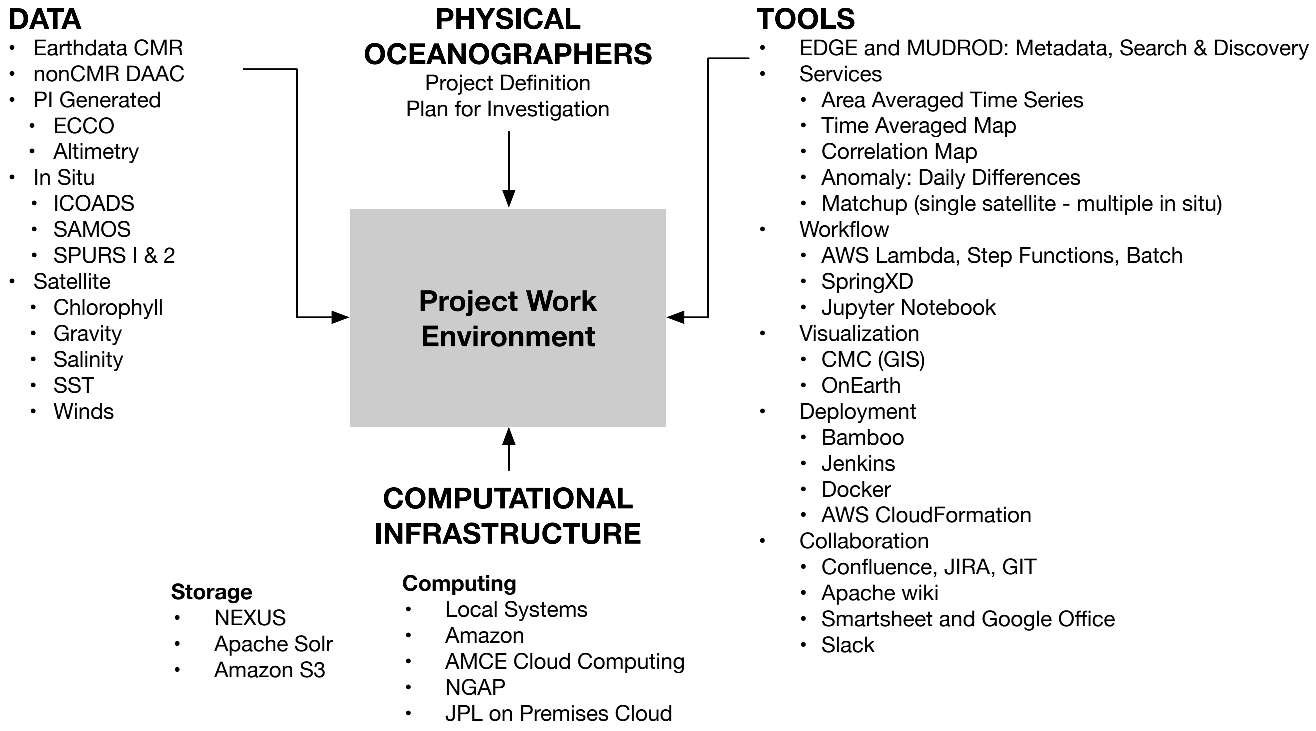
COMPUTATIONAL INFRASTRUCTURE
- Computing
  - Capacity
  - Capability
- Storage
- Communications

Storage
- Data Containers
- Thematic model
- Metadata / Ontology
- Resulting Products
- Published data
- Provenance

Computing
- Local Systems
- High End Computing
- Cloud Computing Capability
- Quantum Computing
- Neuromorphic Computing

Credit: Mike Little, NASA

OceanWorks as an Analytic Center for Ocean Science



DATA

- Earthdata CMR
- nonCMR DAAC
- PI Generated
  - ECCO
  - Altimetry
- In Situ
  - ICOADS
  - SAMOS
  - SPURS I & 2
- Satellite
  - Chlorophyll
  - Gravity
  - Salinity
  - SST
  - Winds

PHYSICAL OCEANOGRAPHERS

Project Definition
Plan for Investigation

Project Work Environment

TOOLS

- EDGE and MUDROD: Metadata, Search & Discovery
- Services
  - Area Averaged Time Series
  - Time Averaged Map
  - Correlation Map
  - Anomaly: Daily Differences
  - Matchup (single satellite - multiple in situ)
- Workflow
  - AWS Lambda, Step Functions, Batch
  - SpringXD
  - Jupyter Notebook
- Visualization
  - CMC (GIS)
  - OnEarth
- Deployment
  - Bamboo
  - Jenkins
  - Docker
  - AWS CloudFormation
- Collaboration
  - Confluence, JIRA, GIT
  - Apache wiki
  - Smartsheet and Google Office
  - Slack

COMPUTATIONAL INFRASTRUCTURE

Storage

- NEXUS
- Apache Solr
- Amazon S3

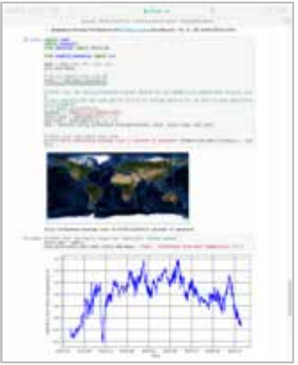
Computing

- Local Systems
- Amazon
- AMCE Cloud Computing
- NGAP
- JPL on Premises Cloud

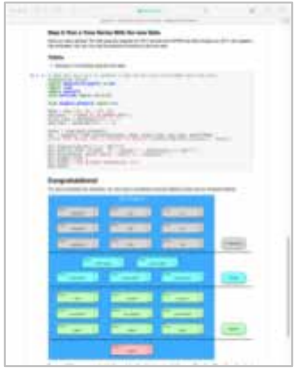
Working with both Science and Informatics Communities

- Established Apache Incubator project
- Develop in the open
- Target Apache top-level project by 2019.
- Public hands-on workshops
- Organize technical sessions at conferences
- Seminars and expert panels
- Lead Editor: 2018 Wiley Book on **Big Earth Data Analytics in Earth, Atmospheric and Ocean Sciences**

Analyze Hurricane Katrina by comparing SST and TRMM time series



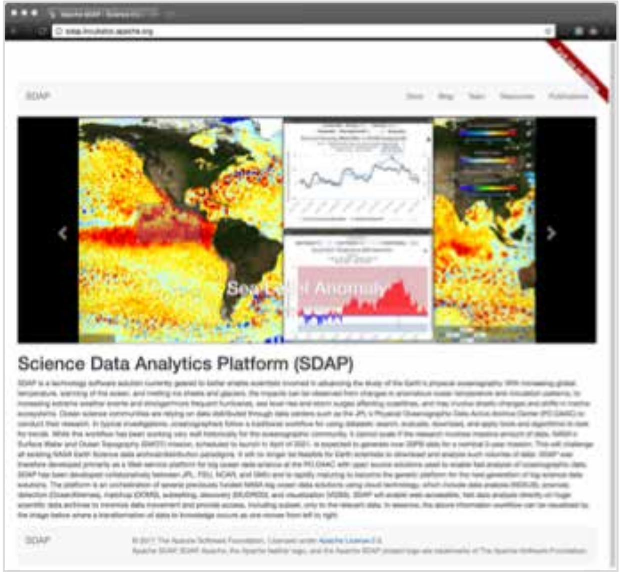
Generate daily difference average
“The Blob” is an oceanographic anomaly



Each participant deployed 3 computing clusters, a total of 24 containers on EC2

Open Source

- Technology sharing through Free and Open Source Software (FOSS)
- Further technology evolution that is restricted by projects / missions
- **Science Data Analytic Platform (SDAP)**, the implementation of **OceanWorks**, in **Apache Incubator**
    - Cloud platform
    - Analyzing satellite and model data
    - In situ data analysis and coordination with satellite measurements
    - Fast data subsetting
    - Mining of user interactions and data to enable discovery and recommendations
    - Streamline deployment through container technology



Science Data Analytics Platform (SDAP)

[illegible]

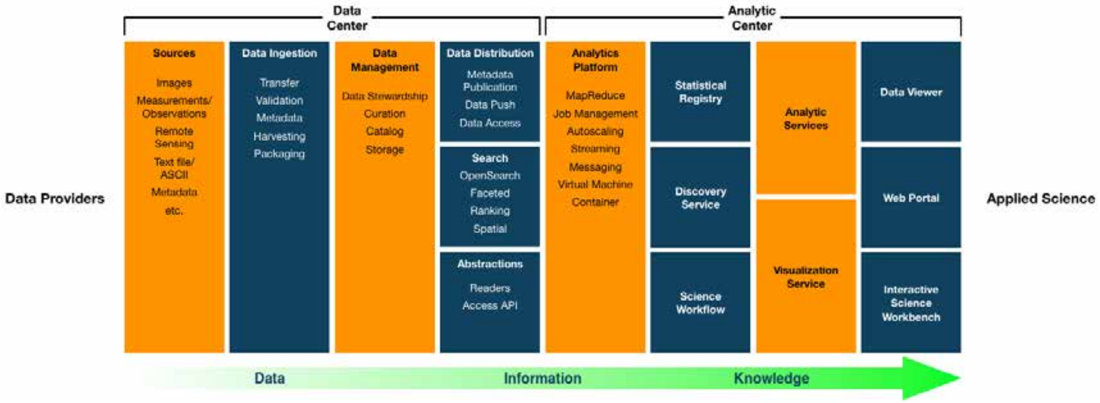
SDAP

http://sdap.incubator.apache.org

In Summary

- Traditional method for scientific research (search, download, local number crunching) is unable to keep up
- Think beyond the archive
- Connected information enables discovery
- Community developed solution through open sourcing
- Thanks to the NASA ESTO/AIST and Sea Level Rise programs, and the NASA ESDIS project
- Investment in data and computational sciences
- Data Centers might want to be in the business of Enabling Science!
- OceanWorks infusion 2018 – 2019
  - Watch for changes to the Sea Level Change Portal
    - Even faster analysis capabilities
    - More variety of measurements – satellites, in situ, and models
    - Event more relevant recommendations
  - NASA’s Physical Oceanography Distributed Active Archive Center (PO.DAAC)

Transforming Data to Knowledge



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Thomas Huang
Jet Propulsion Laboratory
California Institute of Technology

JPL Team

Ed Armstrong, Frank Greguska, Joseph Jacob, Lewis McGibbney, Nga Quach, Vardis Tsontos, and Brian Wilson

Florida State University Team

Shawn Smith, Mark A. Bourassa, Jocelyn Elya

National Center for Atmospheric Research Team

Steve J. Worley, Tom Cram, Zaihua Ji

George Mason University Team

Chaowei (Phil) Yang, Yongyao Jiang, and Yun Li